Exhibit A

GENOME RESEARCH

Large-Scale Gene Expression Data Analysis: A New Challenge to Computational Biologists

Michael Q. Zhang

Genome Res. 1999 9: 681-688
Access the most recent version at doi:10.1101/gr.9.8.681

Perspective

Large-Scale Gene Expression Data Analysis: A New Challenge to Computational Biologists

Michael Q. Zhang

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 USA

The use of high-density DNA arrays to monitor gene expression at a genome-wide scale constitutes a fundamental advance in biology. In particular, the expression pattern of all genes in *Saccharomyces cerevisiae* can be interrogated using microarray analysis where cDNAs are hybridized to an array of each of the ~6000 genes in the yeast genome. In this survey I review three recent experiments related to transcriptional regulation and discuss the great challenge for computational biologists trying to extract functional information from such large-scale gene expression data.

Ever since the theory of genetic regulation of protein synthesis was first worked out almost 40 years ago (Jacob and Monod 1961), biologists have been fascinated by how different genetic programs hard coded in the DNA are involved in the control and regulation of gene expression. This is important because different temporal–spatial gene expression patterns relate directly to developmental control, morphogenesis and cell differentiation, tissue specificity, hormonal communication, or cellular stress responses. Gene expression is largely controlled at the transcriptional level, and transcriptional regulatory elements are located primarily in the upstream promoter region of each gene; however, the lack of quality upstream experimental data has made systematic global investigations very difficult (Zhang 1998a; for review, see Fickett and Hatzigeorgiou 1997). In the past, computational genomics has focused mainly on gene finding (Claverie 1997; Zhang 1997), namely finding the protein coding region and extracting functional information about the protein product. To get equally important functional information about control elements of a gene, one has to analyze functional motifs, most of which occur in noncoding regions (Lavorgna et al. 1998). There have been some excellent works on genomic identification of individual eukaryotic transcriptional factor (TF) binding sites (e.g., Fondrat and Kalogeropoulos 1996; Tronche et al. 1997; Wasserman and Fickett 1998). The most tedious part of these approaches is collecting enough experimentally verified *cis*-acting elements that are shared by a set of coregulated genes.

Since the complete yeast genome has become available, there have been several attempts to computationally identifying upstream activation sequences (UASs) by their clustering properties (Wagner 1997, 1998). This approach suffers from the simple random background (i.e., all oligonucleotides are Poisson distributed) assumption and will certainly miss many TF sites that do not cluster or do not cluster in a simple fashion. Genome-wide monitoring of gene expression has provided a far more effective way of systematically studying coregulated genes and TF sites (for the latest reviews, see "The Chipping Forecast" 1999). I focus on the computational problems of identifying coregulated genes and their promoter elements and refer people to read elsewhere on other interesting and equally challenging issues, such as data requisition/process (Marton et al. 1998), data presentation/management (Ermolaeva et al. 1998), and genetic network dynamics (Altman et al. 1998, 1999).

The power of DNA microarray hybridization on a genome scale was demonstrated early on (Schena et al. 1995; Lockhart et al. 1996; for introduction on the basic concept and protocols, see Stein 1998). For example, in a yeast diauxic shift experiment, five groups of distinct temporal patterns of induction or repression could be recognized visually, as glucose concentration was increased (Fig. 1). The characterized members of each of these groups shared important similarities in their functions, and common regulatory mechanisms could be inferred for most sets of genes with similar expression profiles. When searching for known UAS motifs in each group, many coregulated genes did share common TF sites (DeRisi et al. 1997).

In preparation for regulatory sequence analysis from such expression data, a statistical method was developed (van Helden et al. 1998) based on detection of over-represented oligonucleotides in a target set of upstream sequences over all noncoding sequences from the genome. It was applied to 10 families containing from 5 to 38 genes; 2 of the families were actually built from the DNA microarray expression data of *YAP1* overexpression and *TUP1* deletion (DeRisi et al. 1997). This method was very useful for identifying short core motifs, which is equivalent to the oligonucleotide relative information method (Zhang 1998b) and other methods used in the following experiments.

[1]E-MAIL mzhang@cshl.org; FAX (516) 367-8461.

9:681–688 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

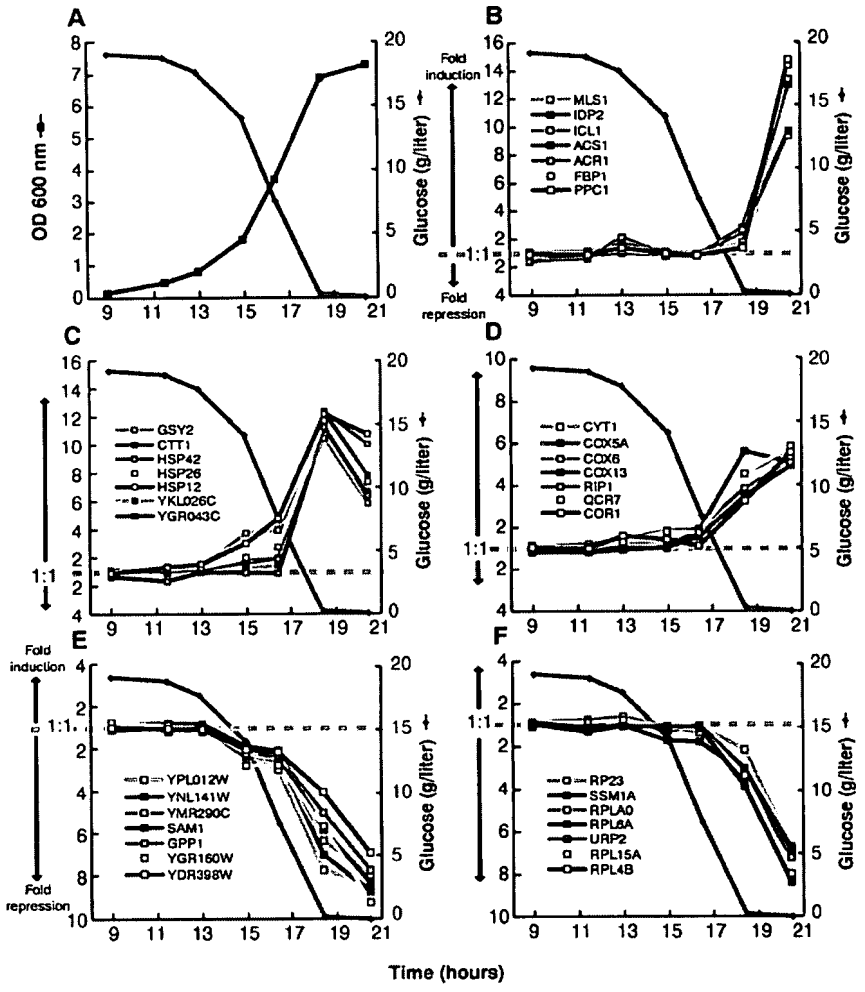

Figure 1 Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. Shown are five examples (*B–D*) of some coregulated genes as glucose concentration (red) dropped during the exponential growing phase (*A*). (Taken from DeRisi et al. 1997).

I now describe real data analyses in the following three sets of recent whole-genome expression experiments. The first was two-point comparisons using oligonucleotide chips, which detected relative mRNA levels before and after nutrient change, heat shock, or mating-type switch (Roth et al. 1998). The second was multipoint (time-course) comparisons also using oligonucleotide chips, which detected mRNA level changes (at different time points after cell cycle release) relative to an arbitrary (but fixed) standard (Cho et al. 1998) for the purpose of identifying cell cycle-regulated genes. The third consisted of both two-point and multipoint comparisons, but using cDNA microarrays, which detected relative mRNA levels (at different time points after cell cycle release) in the synchronous cells relative to the control of asynchronous cells (Spellman et al. 1998) coupled with separate experiments of *CLN3* and *CLB2* induction. In Table 1, the main features in these experiments are listed for easy comparison.

In a two-point experiment (such as in Roth et al. 1998 and in part of Spellman et al. 1998), one measures the relative ratio of mRNA concentration under two different conditions for all genes. After sorting these ratios (one ratio per gene) of mRNA levels, one can identify the most induced or most inhibited genes from the two extreme ends of the sorted list. Some criteria are needed for the gene selection. For example, in the promoter analysis of Roth et al. (1998), the upstream sequences (relative to ATG) of the 10 ORFs were taken from the top and the bottom of the sorted list. Hence a highly induced set, a highly inhibited set, and a combined set were used in searching by AlignACE for common UAS motifs. The rationale for examining the combined set is that a single regulatory motif may serve as either a positive or a negative element depending on its sequence context or environment. AlignACE is a modified version of the Gibbs sampler (Neuwald et al. 1995) and was optimized for finding multiple motifs (via an iterative masking procedure) and for aligning DNA sequences on both strands. It also scores alignments by frequency of occurrence in the intergenic regions of a given genome (for the algorithmic details, see Roth et al. 1998). To suppress false positives, a motif must pass two criteria: (1) exceeding an alignment threshold, and (2) having an occurrence score below 1% (i.e., <1% of genes in the genome may have this motif).

For the galactose versus glucose comparison, UAS_G motif t(T/c)CGG(C/A)(G/c)NNcT(g/c)(T/c)NNcCGG, which is known to regulate galactose utilization genes via the Gal4p/Gal80p activation complex (Lohr et al. 1995) was found successfully, but other expected UASs such as the *Rap1* site, the *Gcr1* site, or the *Mig1* site were not found. For the 39°C vs. 30°C comparison, the heat shock element (HSE) and stress response element (STRE) were not found. Because heat shock is known to

Table 1. Basic Features in Three Genome-Wide Expression Experiments

References	Roth et al. (1998)	Cho et al. (1998)	Spellman et al. (1998)
Goals	Gal-responsive, heat shock and mating type-specific genes and *cis*-acting elements	Cell cycle-regulated genes and *cis*-acting elements	Cell cycle-regulated genes and *cis*-acting elements
Technology	Affymetrix oligo chips	Affymetrix oligo chips	cDNA microarrays
Experiments and synchronization methods	two-point comparisons: (1) galactose vs. glucose; (2) 39°C vs. 30°C; (3) type α vs. type **a**	multipoint comparisons: (1) *cdc28-13* ts mutant;[a] (2) *cdc15-2* ts mutant	two-point comparisons: (1) *cln3*$^+$ vs. *cln3*$^-$; (2) *clb2*$^+$ vs. *clb2*$^-$; Multipoint comparisons: (1) α-factor arrest; (2) elutriation; (3) *cdc15-2* ts mutant
Cluster methods	simple sorting	visual identification of periodic peaks	simple sorting for the two-point data; Fourier transform and Pearson-type correlation for the multipoint
Putative new/total genes identified	(1) 3/9, (2) 6/33, (3) 7/40	~300/416	~700/800
Motif tools	AlignACE	oligonucleotide bias	oligonucleotide information
Database search	TRANSFAC	visual extension TRANSFAC	GibbsDNA TRANSFAC+SCPD
Upstream of ATG	<600 bp	500 bp	700 bp

[a](ts) Temperature sensitive.

have broad effects, including transient cell cycle arrest in G_1 (Rowley et al. 1993), the cell cycle activation (CCA) motif GCGAA(a/g)ttNT(g/c)(a/g)GAA(C/g) of histones was found, but other cell cycle UASs, such as the negative regulatory element of histones (NEG), *Swi4/6* cell cycle box (SCB), *Mlu*I cell cycle box (MCB), and early cell cycle box (ECB) (see below), were not found. For mating-type α versus **a** comparison, all four mating type-specific elements (α2 operator, tCAAtgNcAg; P box, TtCCTAATT(a/g)GgN(c/a)(a/t); pheromone response element (PRE), aTGAAAC; and Q box, tCAAtgNcAg) were found. Some putative motifs were also found, but many were suspected to be false positives. As only one time point (or averaged stationary time points) is taken for each pair of conditions in a two-point comparison experiment, dynamic information is totally lost. It is impossible to separate the primary transcriptional event from the downstream cascades. It would have been much more informative for detecting coregulated genes if measurements had been taken at multiple consecutive time points. (In principle, curve-fitting may result in more robust time series but currently available points in a particular experiment were too limited for smoothing.) This is why I discuss the other two sets of time-course experiments for identifying a complete set of cell cycle-regulated genes and regulatory sequences in yeast.

Figure 2 shows a current model illustrating interactions that determine cell cycle-regulated transcription in yeast (Koch and Nasmyth 1994; McInerny et al. 1997). *Cln3*-associated kinase activates late G_1-specific transcription factors [SBF (SCB binding factor) and MBF (MCB binding factor)] in a cell size-dependent fashion. SBF and MBF mediate the expression of CLN1,2 and CLB5,6 as well as S-phase proteins leading to budding and S-phase entry. CLN1,2 activity allows accumulation of *Clbs* by an unknown mechanism. *Clb1* and *Clb2* activate transcription of G_2-specific genes and thereby autoactivate their own synthesis, possibly via transcription factors Mcm1 and Sff. At the same time, *Clb1,2/cdc28* represses SBF-mediated transcription. Whereas *Clb1,2/cdc28* activates expression of SWI5 and possibly of ACE2 RNAs via *mcm1/Sff*, it keeps the gene products in an inactive state by phosphorylation of the nuclear localization signals. *Clb* proteolysis at the end of mitosis dramatically changes the situation: *Clb*-mediated activation of G_2-specific genes is stopped, and *Swi5* loses its inhibitory phosphorylations, leading to its uptake into the nucleus where it can activate early G_1-specific transcripts. At late M phase, a *mcm1*-related factor binds to ECB (early cell cycle box) and initiates M/G_1-specific activation of CLN3, SWI4, and some DNA replication genes; these gene products have critical roles in promoting the initiation of the next S phase.

Since oscillation of histone mRNAs was discovered (Hereford et al. 1981), 103 cell cycle-regulated messages have been identified using traditional methods,

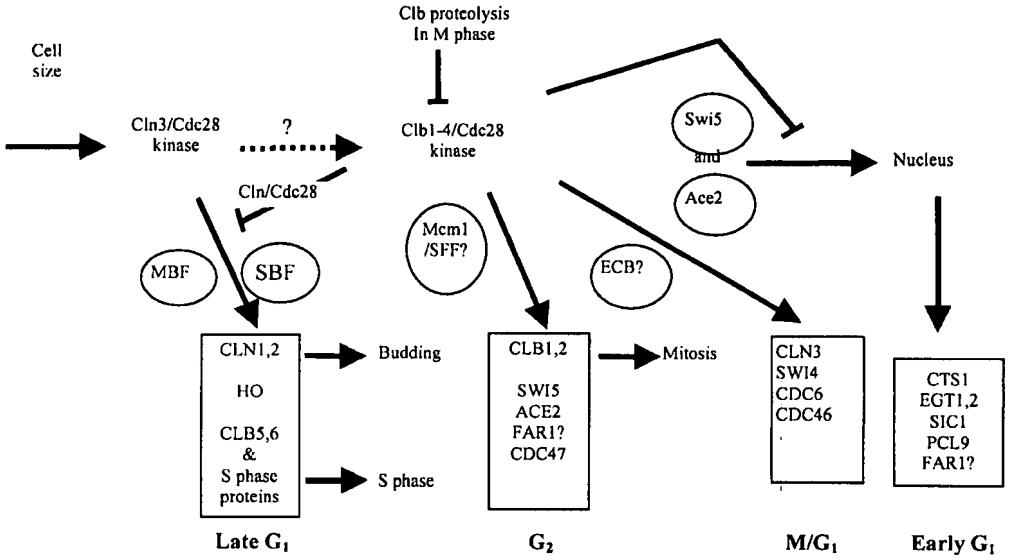

Figure 2 A current model of transcriptional control of the yeast cell cycle.

and it was estimated that some 250 in total might exist (Price et al. 1991). To create a comprehensive catalog of yeast genes whose transcript levels vary periodically with the cell cycle, two independent sets of genome-wide transcriptional experiments have been completed recently (Cho et al. 1998; Spellman et al. 1998).

Cho et al. (1998) used the commercially available oligonucleotide arrays and temperature-sensitive *cdc28* mutant synchronization. After normalization of the expression profiles, cell cycle-dependent periodicity was found for 416 of the ~6200 monitored transcripts. These genes were classified into five groups (early G_1, late G_1, S, G_2, and M) according to their visual peak positions and the consistency with known genes.

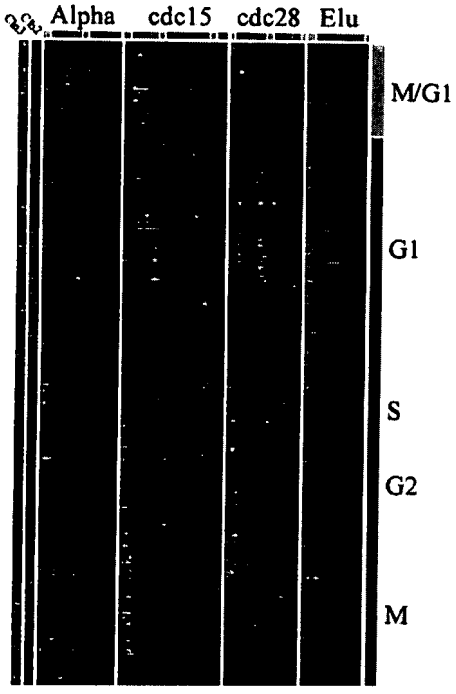

Figure 3 Gene expression patterns of 800 cell cycle-regulated genes are sorted by the phases of the Fourier transform. At each time point, red/green means that synchronous cells have higher/lower expression compared to asynchronous cells. Different phases are color coded on the *top*. The genes having expression levels peak in a specific phase are also coded by the same color scheme at *right*.

Table 2. Example Scores and Oscillation Amplitude for a Collection of Genes

Rank	Score	Gene	Peak to trough
1	15.99	*PIR1*	27.3
9	10.90	*CLN2*	12.1
37	8.78	*CLB1*	9.4
82	6.51	*BUD9*	7.0
177	4.25	*STE3*	12.8
224	3.55	*TUB4*	4.8
255	3.29	*DUN1*	4.2
401	2.37	*CIN8*	5.4
407	2.33	*TUB2*	5.5
585	1.71	*MET1*	3.0
800	1.314	*STP4*	5.9
844	1.28	*SEC8*	4.2
861	1.25	*TUB1*	2.7
1258	0.92	*ANP1*	3.1
1799	0.71	*TUB1*	3.0
2499	0.54	*TUB3*	2.7
2673	0.50	*IME2*	3.5
6054	0.05	*RPS8B*	10.9

See Spellman et al. (1998) for details.

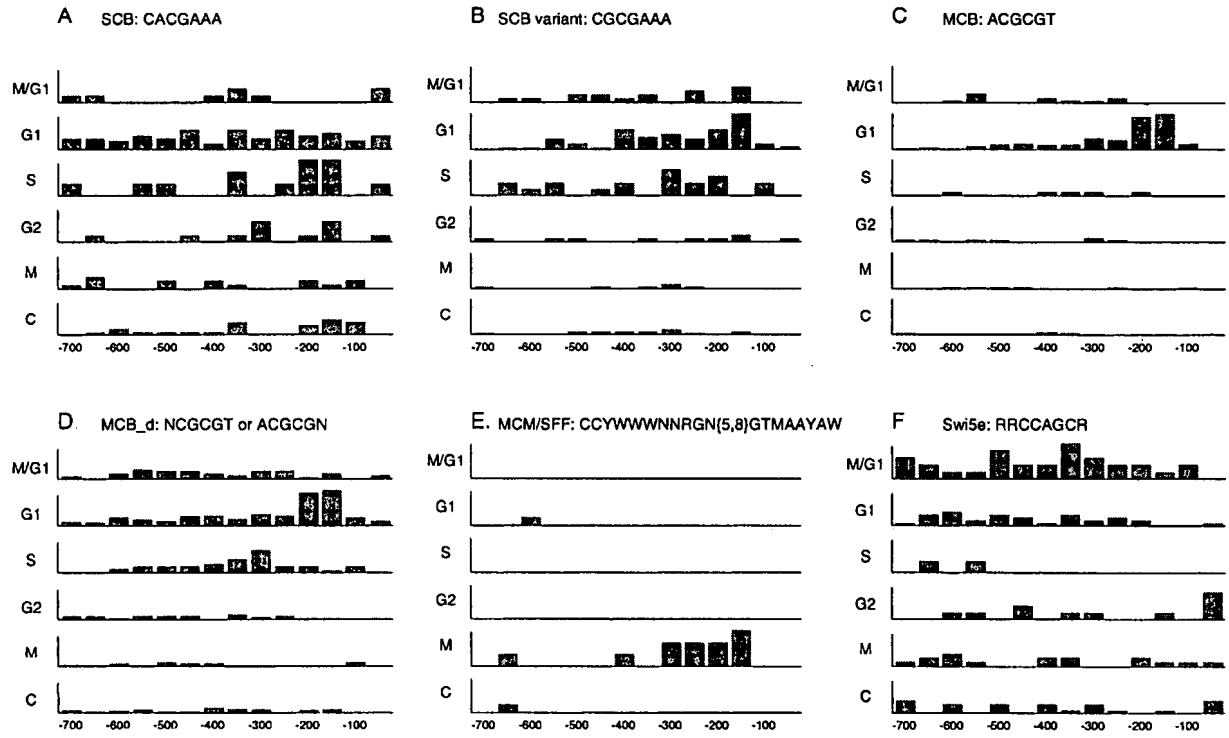

Figure 4 Comparison of some consensus distributions (counts per gene at 50-bp intervals) in the upstream regions of each phase group and control group.

Through matches of known TF sites and visual extension of over-represented short oligonucleotides, a dozen putative motifs (including MCB, SCB, ECB/MCM1) were identified in the promoter regions (Table 2 in Cho et al. 1998).

In contrast, Spellman et al. (1998) used "homemade" cDNA microarrays and samples from yeast cultures synchronized by three independent methods: α-factor arrest, elutriation, and blockage of a *cdc15* temperature-sensitive mutant. Using Fourier transform and Pearson-type correlation algorithms, 800 genes were identified that met an objective minimum criterion for cell cycle regulation. In separate experiments, designed to examine the effects of inducing either the G_1 cyclin Cln3p or the B-type cyclin Clb2p, it was found that mRNA levels of more than half of these 800 genes responded to cyclin induction. These 800 gene expression patterns (sorted by the phases) are shown in Figure 3. The genes were divided into the following five groups: M/G_1 (113 genes), G_1 (300 genes), S (71 genes), G_2 (121 genes), and M (195 genes). Spellman et al. (1998) used the following clustering strategy: Five profiles were built from genes known to be expressed in each class. The averaged peak correlation score (defined as the highest correlation value between the log 2 (ratio) data series for each gene and each profile) of different experiments was used to construct an objective "aggregate CDC (cell cycle-dependent clustering)

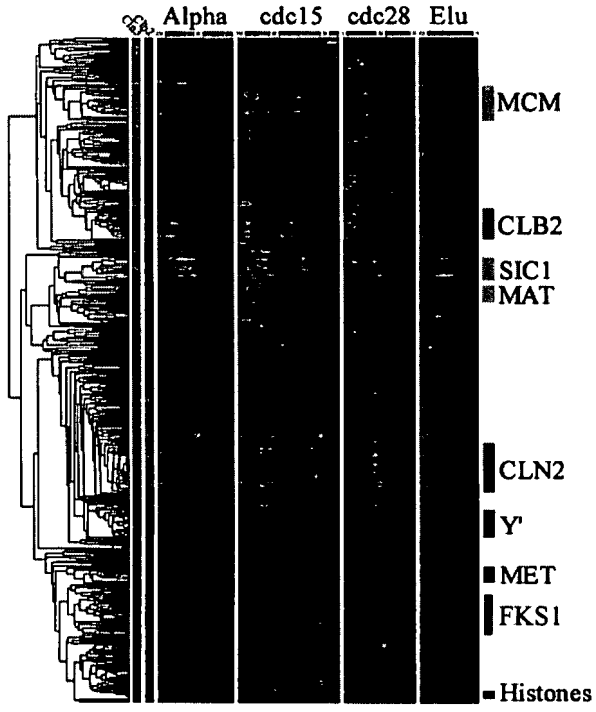

Figure 5 Genes that share similar expression profiles are grouped by correlation clustering (Eisen et al. 1998). The dendragram (*left*) shows the structure of the cluster relationship. Nine clusters for promoter analysis are marked in blue.

Table 3. Potential UAS Motifs Found in Each Cluster

Name	Genes	Group	Motifs	Sites (% genes)	Sites (% genes out of 256 controls)	Cho et al. (1998)
Cln2	58	G_1	MCB:ACGCGT	52 (62)	15 (6)	+
			SCB:CRCGAAA	43 (52)	33 (13)	+
Y′	31	G_1	RAP1:TGCACCW	42 (71)	33 (12)	–
			?:AGCSGCT, etc.	32 (52)	16 (3)	–
Fks1	38	G_1	SCB:CRCGAAA	26 (53)	33 (13)	+
			?:TKCAKCTGCA	4 (11)	3 (1)	–
Histone	9	S	CCA:GcGAArytngrGAACr	19 (100)	0 (0)	–
			NEG:CATTgnGCG	18 (89)	1 (0)	–
			SCB:CGCGAAA	7 (56)	14 (5)	+
Met	20	S/G_2	Cbf1/Met/Met28:TCACGTG	20 (60)	17 (5)	–
			Met31/Met32:AAAnTGTGG	14 (55)	12 (5)	–
Clb2	36	G_2/M	Mcm1(P-box):TTWCCyaawnnGGwAA	55 (64)	1 (0)	+
			Mcm1(P-box):+Sff:(P)n_{2-4}RTaAAYAA	19 (47)	0 (0)	–
Mcm	38	M/G_1	ECB:TTTCCcaATngGGAAA	73 (79)	1 (0)	+
			?:AAAGAAAA	26 (53)	20 (8)	–
Sic1	27	M/G_1	SWI5:RRCCAGCR	23 (48)	23 (9)	–
			?:GCSCRGC	12 (41)	31 (11)	–
Mat	13	M/G_1	Ste12(PRE):TGAAACA	10 (54)	48 (18)	–
			P′+Q:tTTCCTaaTTrGknnnTCAATG	8 (46)	0 (0)	–
			?:WnAnnAGCCAnnnnWWnMAAAnA	6 (46)	2 (1)	–

(?) An unknown motif. Motif site counts (percent of genes containing the motif) in each cluster and in the control are also shown. (+ or –) The motif was found or not found in Cho et al. (1998). As Y′ is full of repeats, there are many "motifs" that look significant on pure statistical ground. All sites were counted on both strands in the (–700, –1) region, except MCB:ACGCGT was counted on one strand and histone motifs were counted only in the commonly shared promoter regions.

score." Genes were ranked by their aggregate CDC scores, and the list was examined to determine a threshold that was exceeded by 91% of known cell cycle-regulated genes. Altogether, 800 genes met or exceeded the threshold. Clustering of randomly shuffled data indicated that the false-positive rate should be <10%. Table 2 provides some examples of the kinds of scores obtained for several genes (including specific examples that are and are not cell cycle regulated).

Using a newly developed *Saccharomyces cerevisiae* promoter database and analysis tools (Zhu and Zhang 1999), the 700-bp upstream regions of the five phase groups were analyzed further.

Spellman et al. (1998) first computed the relative pentamer information (an oligonucleotide bias measure) of each phase group versus the control group of non-cell-cycle genes (Zhang 1998b). They then tried to extend the informative oligomers or to find other longer motifs by using GibbsDNA (Z. Ioschikhes and M.Q. Zhang, unpubl.), which is another modified version of the Gibbs sampler and includes features, such as double strand, palindrome symmetry, distance constraint and submotif inclusion/exclusion. As Gibbs sampling is a stochastic process, a sufficient number of runs had to be carried out for each data set with various parameters, and the motifs that had higher maximum aposteri probability (MAP) values were selected. Once motifs were established for a group, their predictive value was tested by searching for the motif consensus (with specified mismatches) in the promoter regions of all groups, as well as for the control group. Figure 4 shows the selectivity of some consensus motifs with respect to different regulatory groups and positions.

Because the cutoffs for different phase groups were somewhat arbitrary, to search for better coregulated gene clusters, the Peason-type correlation clustering algorithm (Eisen et al. 1998) was used to identify nine clusters [data from Cho et al. (1998) was also included during clustering for completeness]. The dendragram of these clustered expression profiles is shown in Figure 5. UAS motifs for each of these clusters were calculated, and results are shown in Table 3. As the statistics show, many of these motifs contain information predictive of cell cycle regulation (see Fig. 4). A full description and complete data sets are available at http://cellcycle-www.stanford.edu and at http://www.cshl.org/mzhanglab.

It is clear that multiple time points are more useful and better for clustering and promoter analysis. The most obvious difference between the results of Spellman et al. (1998) and Cho et al. (1998) is the number of cell cycle-regulated genes and promoter elements

identified. With a manual decision process, Cho et al. found 421 genes to be cell cycle regulated. A set of 800 genes found by Spellman et al. includes 304 of these, but the other 117 do not appear significantly cell cycle regulated in their experiments. The set of 800 genes therefore contains 496 genes not identified by Cho et al. The main cell cycle control promoter element SWI5 site (among some others) was not identified in the analysis of Cho et al. because (according to Cho et al.) "*Swi5* do not have a highly conserved binding sequence, making it difficult to accurately search genomic sequence for possible sites of action." Here, the SWI5 site was a good example for which the factor was well known generally, but the site had not been well characterized experimentally. Another advantage of the analysis of Spellman et al. was the diversity of experiments, which allowed them to distinguish cell cycle regulation from confounding patterns such as those caused by the heat shock response when a culture is shifted from one temperature to another. Although transcriptional cascade could be better resolved by adding more time points, there are certainly technical limits. Because differential stability could also affect the transcript level as well as transcription rate, a systematic detection of the turnover rate for each transcript would be also crucial for more accurate global picture.

In summary, with the help of genome-wide expression techniques, it is possible to identify coregulated genes by clustering analysis. Furthermore, by combination of over-represented oligonucleotide analysis and multiple-sequence alignment programs, it is also possible to identify upstream regulatory motifs commonly shared by coregulated genes. Good clustering is better than sophisticated motif-search algorithms. It would be highly desirable if one could combine motif and cluster analyses, as good clustering can facilitate motif identification, and, conversely, conserved motifs (or any other functional information related to the sequences) can help to improve clustering. We ought to work toward a self-consistent iteration process (clustering coregulated genes → detecting common motifs) as used in all scientific inference (functional groups → conserved structures). Although I have only discussed promoter motif detection in the context of array data analysis, it can be equally applied to a similar search of 3′-untranslated regions (3′ UTRs). As the 3′ UTR is often important for message stability and transport, identification of conserved motifs in this region may be also be instructive for gene regulation.

ACKNOWLEDGMENTS

I thank all my collaborators: P.T. Spellman, G. Sherlock, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Z. Ioschikhes, and J. Zhu. I am grateful to F.P. Roth for providing the preprint. I also thank R.W. Davis for helpful discussions during a visit to the Stanford Genome Center and T.G. Wolfsberg for useful references. I further thank Lincoln Stein for reading and commenting on the manuscript. My laboratory is supported by National Institutes of Health/National Human Genome Research Institute (NIH/NHGRI; HG01696), Cold Spring Harbor Laboratory (CSHL) Association, and Merck Genome Research Institute.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

NOTE

Recently, an interesting new clustering algorithm GENECLUSTER based on self-organizing maps has been developed (Tamayo et al. 1999); it was used to recluster the data of Cho et al. (1998) and found essentially similar results. Unfortunately, there was no regulatory sequence analysis.

REFERENCES

Altman, R.B., A.K. Dunker, L. Hunter, and T.E. Klein eds. 1998. Gene expression and genetic networks, session in *Biocomputing: Proceedings of the 1998 Pacific Symposium.* World Scientific, Singapore.

———. 1999. Gene expression and genetic networks, session in *Biocomputing: Proceedings of the 1999 Pacific Symposium.* World Scientific, Singapore.

Cho, R.J., M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.

"The Chipping Forecast." 1999, a special supplement to *Nat. Genet.* Vol. 21, January.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Ermolaeva, O., M. Rastogi, K.D. Pruitt, G.D. Schuler, M.L. Bittner, Y. Chen, R. Simon, P. Meltzer, J.M. Trent, and M.S. Boguski. 1998. Data management and analysis for gene expression arrays. *Nat. Genet.* **20:** 19–23.

Fickett, J.W. and A.G. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Res.* **7:** 861–878.

Fondrat, C. and A. Kalogeropoulos. 1996. Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: Application to chromosome III. *Comput. Appl. Biosci.* **12:** 363–374.

Hereford, L.M., M.A. Osley, T.R.D. Ludwig, and C.S. McLaughlin. 1981. Cell-cycle regulation of yeast histone mRNAs. *Cell* **24:** 367–375.

Jacob, F. and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3:** 318–356.

Koch, C. and K. Nasmyth. 1994. Cell cycle regulated transcription in yeast. *Curr. Opin. Cell Biol.* **6:** 451–459.

Lavorgna, G., E. Boncinelli, A. Wagner, and T. Werner. 1998. Detection of potential target genes *in silico*? *Trends Genet.* **14:** 375–376.

Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*

14: 1675–1680.

Lohr, D., P. Venkov, and J. Zlatanova. 1995. Transcriptional regulation in the yeast GAL gene family: A complex genetic network. *FASEB J.* **9:** 777–787.

Marton, M.J., J.L. DeRisi, H.A. Bennett, V.R. Iyer, M.R. Meyer, C.J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai et al. 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* **4:** 1293–1307.

McInerny, C.J., J.F. Partridge, G.E. Mikesell, D.P. Creemer, and L. Breeden. 1997. A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G_1-specific transcription. *Genes & Dev.* **11:** 1277–1288.

Neuwald, A.F., J.S. Liu, and C.E. Lawrence. 1995. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.* **4:** 1618–1632.

Price, C., K. Nasmyth, and T. Schuster. 1991. A general approach to the isolation of cell cycle regulated genes in the budding yeast, Saccharomyces cereviase. *J. Mol. Biol.* **218:** 543–556.

Roth, F.P., J.D. Hughes, P.W. Estep, and G.M. Church. 1998. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.

Rowley, A., G.C. Johnston, B. Butler, M. Werner-Washburne, and R.A. Singer. 1993. Heat shock-mediated cell cycle blockage and G1 cyclin expression in the yeast Saccharomyces cerevisiae. *Mol. Cell. Biol.* **13:** 1034–1041.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.

Spellman, P.T., G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9:** 3273–3297.

Stein, L.D. 1998. Genetic Analysis on DNA microarrays. In *Current Protocols in Human Genetics* (ed. N.C. Dracopoli, J.L. Haines, B.R. Korf, D.T. Moir, C.C. Morton, C.E. Seidman, J.G. Seidman, and D.R. Smith), pp. 7.9.1–7.9.8. John Wiley and Sons, New York, NY.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **96:** 2907–2912.

Tronche, F., F. Ringeisen, M. Blumenfeld, M. Yaniv, and M. Pontoglio. 1997. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266:** 231–245.

van Helden, J., B. Andre, and J. Collado-Vides. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281:** 827–842.

Wagner, A. 1997. A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.* **25:** 3594–3604.

———. 1998. Distribution of transcription factor binding sites in the yeast genome suggests abundance of coordinately regulated genes. *Genomics* (in press).

Wasserman, W. and J.W. Fickett. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.

Zhang, M.Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci.* **94:** 565–568.

———. 1998a. Identification of human gene core promoters in silico. *Genome Res.* **8:** 319–326.

———. 1998b. Promoter analysis of co-regulated genes in the yeast. *Comput. Chem.* **23:** 233–250.

Zhu, J. and M.Q. Zhang. 1999. SCPD: a promoter database of yeast *Saccharomyces cerevisiae*. *Bioinformatics* (in press).